**Content**

[1. TODO: To successfully pass the exam 3](#_Toc21876168)

[2. SageMaker built-in Algorithms -overview 4](#_Toc21876169)

[1. Common parameters 4](#_Toc21876170)

[2. What is the meaning of “downstream” processing? 4](#_Toc21876171)

[3. What is the difference between Classification and Regression? 4](#_Toc21876172)

[2.3.1. Function Approximation 5](#_Toc21876173)

[2.3.2. Classification Predictive Modelling 5](#_Toc21876174)

[2.3.3. Regression Predictive Modelling 5](#_Toc21876175)

[2.3.4. Classification v Regression 5](#_Toc21876176)

[2.3.5. Convert between Classification and Regression 6](#_Toc21876177)

[4. Ensemble Learning 6](#_Toc21876178)

[5. CART 6](#_Toc21876179)

[6. Crowdsourcing 6](#_Toc21876180)

[3. Algorithms one by one 7](#_Toc21876181)

[1. BlazingText 7](#_Toc21876182)

[2. DeepAR Forecasting 7](#_Toc21876183)

[3. Factorization Machines 7](#_Toc21876184)

[4. Image Classification 8](#_Toc21876185)

[5. IP Insights 8](#_Toc21876186)

[6. k-means 8](#_Toc21876187)

[7. k-NN (K-Nearest Neighbors) 8](#_Toc21876188)

[8. LDA – Latent Dirichlet Allocation 8](#_Toc21876189)

[9. Neural Topic Model 8](#_Toc21876190)

[10. Object2Vec 8](#_Toc21876191)

[11. Object Detection 9](#_Toc21876192)

[12. PCA 9](#_Toc21876193)

[13. RFC 9](#_Toc21876194)

[14. Semantic Segmentation 9](#_Toc21876195)

[15. Seq2Seq modelling 9](#_Toc21876196)

[16. XGBoost 9](#_Toc21876197)

[17. Linear Learner 9](#_Toc21876198)

[18. RNN – Recurrent Neural Network 9](#_Toc21876199)

1. TODO: To successfully pass the exam
2. LA videos
3. Community slack pinned staff.
4. **SageMaker docs (900 pages ?!) - algorithm chapter.**
5. Sagemaker FAQS
6. Example questions w links
7. Practice exam
8. Copy practice exam question to google and fins similar
9. Youtube ML vid list

* <https://www.youtube.com/playlist?list=PLhr1KZpdzukeKUChZG8SdYipk7INwphQm>
* <https://www.youtube.com/playlist?list=PLhr1KZpdzukdExxUr05-WIpdVYPWjgi43>

1. SageMaker built-in Algorithms -overview

Model and Inferences:

A machine learning algorithm uses example data to **create a generalized solution** (a **model**) that addresses the business question you are trying to answer. After you create a model using example data, you can use it to **answer the same business question for a new set of data**. This is also referred to as obtaining **inferences**.

* 1. Common parameters

Algorithms that are **parallelizable** can be deployed on multiple compute instances for distributed training. For the **Training Image and Inference Image Registry Path** column, use the :1 version tag to ensure that you are using a stable version of the algorithm. You can reliably host a model trained using an image with the :1 tag on an inference image that has the :1 tag. Using the :latest tag in the registry path provides you with the most up-to-date version of the algorithm, but might cause problems with backward compatibility. **Avoid using the :latest tag for production purposes.**

**Training Image and Inference Image Registry Path**

Use the paths and training input mode as follows:

* To create a **training job** (with a request to the CreateTrainingJob API), specify the Docker Registry path and the training input mode for the training image. You create a training job to train a model using a specific dataset.
* To **create a model** (with a CreateModel request), specify the Docker Registry path for the inference image. Amazon SageMaker launches machine learning compute instances that are based on the endpoint configuration and deploys the model, which includes the artifacts (the result of model training).

XGBoost implements an open-source algorithm that has been optimized for CPU computation.

* 1. What is the meaning of “downstream” processing?

"Downstream tasks is what the field calls those supervised-learning tasks that utilize a pre-trained model or component"

* 1. What is the difference between Classification and Regression?

*Fundamentally, classification is about predicting a label and regression is about predicting a quantity.*

* **Predictive modeling** is about the problem of learning a mapping function from inputs to outputs called function approximation.
* C**lassification** is the problem of predicting a discrete class label output for an example.
* **Regression** is the problem of predicting a continuous quantity output for an example.
  + 1. Function Approximation

Predictive modeling is the problem of developing a model using historical data to make a prediction on new data where we do not have the answer. Predictive modeling can be described as the mathematical problem of approximating a mapping function (f) from input variables (X) to output variables (y). This is called the problem of function approximation. The job of the modeling algorithm is to find the best mapping function we can give the time and resources available

* + 1. Classification Predictive Modelling

Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

* A classification problem requires that examples be classified into one of two or more classes.
* A classification can have real-valued or discrete input variables.
* A problem with two classes is often called a two-class or binary classification problem.
* A problem with more than two classes is often called a multi-class classification problem.
* A problem where an example is assigned multiple classes is called a multi-label classification problem.
  + 1. Regression Predictive Modelling

Regression predictive modeling is the task of approximating a mapping function (f) from input variables (X) to a continuous output variable (y).

* + 1. Classification v Regression
* Classification is the task of predicting a discrete class label.
* Regression is the task of predicting a continuous quantity.
* A classification algorithm may predict a continuous value, but the continuous value is in the form of a probability for a class label.
* A regression algorithm may predict a discrete value, but the discrete value in the form of an integer quantity.
* Classification predictions can be evaluated using accuracy, whereas regression predictions cannot.
* Regression predictions can be evaluated using root mean squared error, whereas classification predictions cannot.
  + 1. Convert between Classification and Regression

Class 0: $0 to $49

Class 1: $50 to $100

* 1. Ensemble Learning

In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Unlike a statistical ensemble in statistical mechanics, which is usually infinite, a machine learning ensemble consists of only a concrete finite set of alternative models, but typically allows for much more flexible structure to exist among those alternatives.

* 1. CART

Entropy gain …

* 1. Crowdsourcing

Crowdsourcing is a sourcing model in which individuals or organizations **obtain goods and services, including ideas and finances, from a large, relatively open and often rapidly-evolving group of internet users**; it divides work between participants to achieve a cumulative result. The word crowdsourcing itself is a portmanteau of crowd and outsourcing, and was coined in 2006. As a mode of sourcing, crowdsourcing existed prior to the digital age (i.e. "offline").

1. Algorithms one by one
   1. BlazingText

sentiment analysis, named entity recognition, machine translation, etc. Text classification is an important task for applications that perform web searches, information retrieval, ranking, and document classification.

The hyperparameters for the BlazingText algorithm depend on which mode you use: Word2Vec (unsupervised) and Text Classification (supervised).

* 1. DeepAR Forecasting

The Amazon SageMaker DeepAR forecasting algorithm is a supervised learning algorithm for forecasting scalar (one-dimensional) time series using recurrent neural networks (RNN). Classical forecasting methods, such as autoregressive integrated moving average (ARIMA) or exponential smoothing (ETS), fit a single model to each individual time series. They then use that model to extrapolate the time series into the future.

Predict price, labor needs.

**RMSE – root mean square error**

* 1. Factorization Machines

A factorization machine is a general-purpose supervised learning algorithm that you can use for both classification and regression tasks. It is an extension of a linear model that is designed to capture interactions between features within high dimensional sparse datasets economically. For example, in a click prediction system, the factorization machine model can capture click rate patterns observed when ads from a certain ad-category are placed on pages from a certain page-category. Factorization machines are a good choice for tasks dealing with high dimensional sparse datasets, such as click prediction and item recommendation.

The Amazon SageMaker implementation of factorization machines considers only **pair-wise** (2nd order) **interactions** **between** **features**.

Are a new model class that combines the advantages of Support Vector Machines (SVM) with factorization models Like SVMs, FMs are a general predictor working with any real valued feature vector. In contrast to SVMs, FMs model all interactions between variables using factorized parameters. Thus they are able to estimate interactions even in problems with huge sparsity (like recommender systems) where SVMs fail.

* 1. Image Classification
  2. IP Insights
  3. k-means

K-means is an unsupervised learning algorithm. It attempts to find discrete groupings within data, where members of a group are as similar as possible to one another and as different as possible from members of other groups. You define the attributes that you want the algorithm to use to determine similarity. Clustering.

* 1. k-NN (K-Nearest Neighbors)

K-nearest neighbor is a subset of supervised learning classification (or regression) algorithms (it takes a bunch of labeled points and uses them to learn how to label other points). It is supervised because you are trying to classify a point based on the known classification of other points. In contrast, K-means is a subset of unsupervised learning clustering algorithms (it takes a bunch of unlabeled points and tries to group them into clusters). It is unsupervised because the points have no external classification.

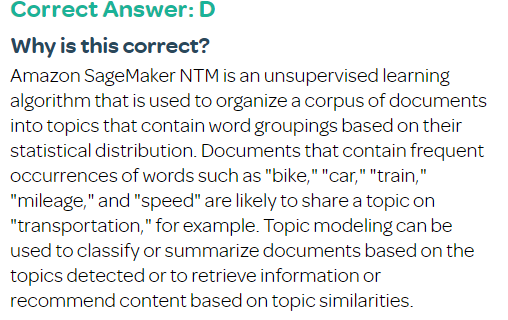
The k-nearest neighbors algorithm (k-NN) is a supervised, non-parametric method used for classification and regression. The input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression.

- In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

- In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors

* 1. LDA – Latent Dirichlet Allocation

Unsupervised. Most commonly used to discover a user-specified number of topic shared by documents within a text corpus.

* 1. Neural Topic Model

Unsupervised. Used to organize a corpus of documents into topics that contain word groupings based on their statistical distribution. Topic modelling can be used to classify or summarize documents based on the topic detected or to retrieve information or recommend content based on topic similarities.

* 1. Object2Vec

Object2Vec generalizes the well-known Word2Vec embedding technique for words that is optimized in the Amazon SageMaker BlazingText Algorithm

* 1. Object Detection
  2. PCA
  3. RFC

Detecting anomalous data points within a data set.

* 1. Semantic Segmentation

Provides a fine-grained, pixel-level approach to developing computer vision applications. It tags every pixel in an image with a class label from a predefined set of classes. Tagging is fundamental for understanding scenes, which is critical to an increasing number of computer vision applications, such as self-driving vehicles, medical imaging diagnostics, and robot sensing.

* 1. Seq2Seq modelling
  2. XGBoost

Supervised, quantitative analysis (regression). Ensemble Learning. Ranking, finding relevant scores, uses CART (Classification and Regression Trees).

* 1. Linear Learner

Supervised, both classification and regression problems.

The best model optimizes either of the following:

* Continuous objectives, such as mean square error, cross entropy loss, and absolute error.
* Discrete objectives suited for classification, such as F1 measure, precision, recall, or accuracy.
  1. RNN – Recurrent Neural Network

Can be used for sentiment analysis.